EfficientMedSAM: Accelerating Medical Image Segmentation via Neural Architecture Search and Knowledge Distillation

Arnold Caleb Asiimwe
Columbia University
New York, USA
aa4870@columbia.edu

William Das
Columbia University
New York, USA
whd2108@columbia.edu

Hadjer Benmeziane
IBM Research Europe
Rüschlikon, Switzerland
Hadjer.Benmeziane1@ibm.com

Kaoutar El Maghraoui
IBM T. J. Watson Research Center
Yorktown Heights, NY, USA
kelmaghr@us.ibm.com

***Abstract*—Medical image segmentation is crucial for precise diagnosis, treatment planning, and disease monitoring in clinical practice. Despite the advancements in segmentation models inspired by the Segment Anything Model (SAM) architecture, their real-time application is limited due to computational inefficiencies arising from transformer-based architectures. We introduce EfficientMedSAM, a suite of high-speed, memory-efficient foundation models for universal medical image segmentation. Our approach leverages differentiable neural architecture search (NAS) to explore a novel search space emphasizing efficient operations over traditional attention mechanisms. The generated candidate architectures are further refined through knowledge distillation from larger MedSAM models and evaluated on the Kvasir dataset of endoscopic images. EfficientMedSAM achieves competitive mean Average Precision (mAP) while substantially reducing Multiply-Accumulate operations (MACs) and model parameters, enhancing throughput. We integrate a knowledge distillation (KD) pipeline that transfers knowledge from logits and attention maps, using saliency maps as proxies for attention map-based distillation. Our findings establish a proof of concept for large-scale distributed training on the SA-Med2D-20M dataset, paving the way for real-time medical image segmentation advancements.**

***Index Terms*—Medical Image Segmentation, Neural Architecture Search, Knowledge Distillation, Edge Computing**

I. Introduction

Medical image segmentation is a critical task in computer vision that aims to identify and delineate regions of interest, such as organs, lesions, or tumors, in medical images [14]. It plays a vital role in various clinical applications, including diagnosis, treatment planning, and disease monitoring [28, 25]. Recently, the introduction of the Segment Anything Model (SAM) [11] has revolutionized the field of image segmentation by providing a foundation model capable of segmenting any object in an image with minimal input. Inspired by SAM, several works have proposed medical image segmentation models that leverage transformer architectures to achieve state-of-the-art performance [38, 26, 18].

However, the superior performance of these transformer-based models comes at the cost of high computational complexity and memory requirements, making them unsuitable for real-time clinical applications [5, 31]. The heavy reliance on multi-head self-attention (MHSA) modules, which involve memory-inefficient operations such as tensor reshaping and element-wise functions, is a major bottleneck in terms of inference speed [33, 9]. Moreover, the redundancy in attention maps across different heads leads to unnecessary computational overhead [19, 32]. These challenges hinder the deployment of transformer-based image segmentation models in resource-constrained environments and limit their practical utility.

To address these limitations, we propose **EfficientMedSAM**, a family of high-speed and memory-efficient foundation models for universal medical image segmentation. Our approach leverages differentiable neural architecture search [16, 39] to discover efficient alternatives to attention mechanisms in the transformer architecture. By exploring a search space that includes memory-efficient operations, we identify candidate architectures that strike a balance between computational efficiency and segmentation performance. We further enhance these architectures by distilling knowledge from larger MedSAM models [7, 24], allowing them to inherit the strong representational power of the teacher models while maintaining a lightweight structure.

We evaluate the proposed EfficientMedSAM models on the Kvasir-SEG dataset [22], a subset of the Kvasir dataset specifically designed for image segmentation tasks. The Kvasir-SEG dataset consists of 1,000 endoscopic images with corresponding segmentation masks. Our models demonstrate their effectiveness by achieving competitive mean Average Precision (mAP) scores while significantly reducing the number of Multiply-Accumulate operations (MACs) and model parameters. EfficientMedSAM models maintain high segmentation accuracy while substantially improving inference speed and memory efficiency compared to existing transformer-based approaches.

The main contributions of this work are as follows: (1) We propose EfficientMedSAM, a family of high-speed and memory-efficient foundation models for universal medical image segmentation, discovered through differentiable neural architecture search and knowledge distillation. (2) We propose a search space that includes memory-efficient operations as alternatives to attention mechanisms, enabling the discovery of

architectures that strike a balance between computational efficiency and segmentation performance. (3) We demonstrate the effectiveness of the EfficientMedSAM models on the Kvasir dataset, achieving competitive mAP scores while significantly reducing MACs, model parameters, and increasing throughput compared to existing transformer-based approaches. (4) We provide a proof of concept for large-scale distributed training of efficient medical image segmentation models, setting the stage for future work on more extensive datasets like SA-Med2D-20M [36].

The remainder of the paper is organized as follows: Section II reviews related work on medical image segmentation and efficient transformer architectures. Section III describes the proposed EfficientMedSAM framework, including the neural architecture search and knowledge distillation components. Section IV presents the experimental setup and results on the Kvasir dataset. Finally, Section V concludes the paper and discusses future directions.

II. Related Work

A. Medical Image Segmentation

Medical image segmentation has been an active area of research for decades, with numerous methods proposed to tackle the challenges of accurately delineating regions of interest in medical images [14]. Traditional approaches, such as thresholding, region growing, and level sets, rely on handcrafted features and often struggle with the variability and complexity of medical images [21]. With the advent of deep learning, convolutional neural networks (CNNs) have become the dominant approach for medical image segmentation, achieving state-of-the-art performance on various modalities and anatomical structures [25, 20].

Transformer-based transformer-based architectures have shown impressive results on a wide range of computer vision tasks, including medical image segmentation. The Segment Anything Model (SAM) [11] has revolutionized the field of image segmentation by providing a foundation model capable of segmenting any object in an image with minimal input.

Inspired by SAM, several works have proposed medical image segmentation models that leverage transformer architectures to achieve state-of-the-art performance. MedSAM [18], a transformer-based model specifically designed for medical image segmentation, has demonstrated superior performance compared to CNN-based approaches. However, the computational complexity and memory requirements of MedSAM and other transformer-based models limit their applicability in resource-constrained environments and real-time clinical settings.

B. Efficient Neural Architectures

The increasing complexity of deep learning models, coupled with the growing demand for real-time applications, has motivated the development of efficient neural architectures [1]. In the context of CNNs, architectures like MobileNet [8], ShuffleNet [37], and EfficientNet [30] have been proposed to reduce the computational cost and memory footprint while maintaining high performance. These architectures employ techniques such as depth-wise separable convolutions, channel shuffling, and neural architecture search to strike a balance between efficiency and accuracy.

However, the trend of increasing model complexity and computational requirements has been at odds with the slower growth of memory and computation resources in edge devices. This discrepancy highlights the need for efficient neural architectures that can operate within the constraints of resource-limited environments, such as mobile devices, augmented reality headsets, or embedded systems.

With regards to transformers, efforts have been made to design efficient architectures that reduce the computational burden while retaining the benefits of self-attention mechanisms. Methods like Sparse Transformers [4], Linformer [33], and Reformer [12] propose techniques such as sparse attention, low-rank approximations, and locality-sensitive hashing to improve the efficiency of transformers. However, these approaches often come with a trade-off in terms of performance, and their applicability to medical image segmentation tasks remains to be explored.

C. Neural Architecture Search

Neural architecture search (NAS) has emerged as a powerful tool to automatically discover efficient neural network architectures [6]. NAS algorithms explore a search space of architectural components and connections, optimizing for specific objectives such as accuracy or efficiency. Differentiable NAS methods have gained popularity due to their ability to efficiently search large spaces by relaxing the discrete architecture representation into a continuous one, allowing for gradient-based optimization [16, 35]. In the context of medical image segmentation, NAS has been applied to discover efficient CNN architectures [34, 10]. However, the application of NAS to transformer-based models for medical image segmentation remains an open research question.

D. Knowledge Distillation

Knowledge distillation has been proposed as a technique to transfer knowledge from a large, complex teacher model to a smaller, more efficient student model [7]. By mimicking the behavior of the teacher model, the student model can achieve comparable performance while being more lightweight and computationally efficient. Knowledge distillation has been successfully applied to various computer vision tasks, including image classification, object detection, and semantic segmentation [2, 17].

In the context of medical image segmentation, knowledge distillation has been explored to improve the efficiency of CNN-based models [23, 3]. However, the application of knowledge distillation to transformer-based models for medical image segmentation has not been extensively studied. By leveraging knowledge distillation, we aim to transfer the knowledge learned by larger, more complex MedSAM models to our efficient EfficientMedSAM architectures, enabling them

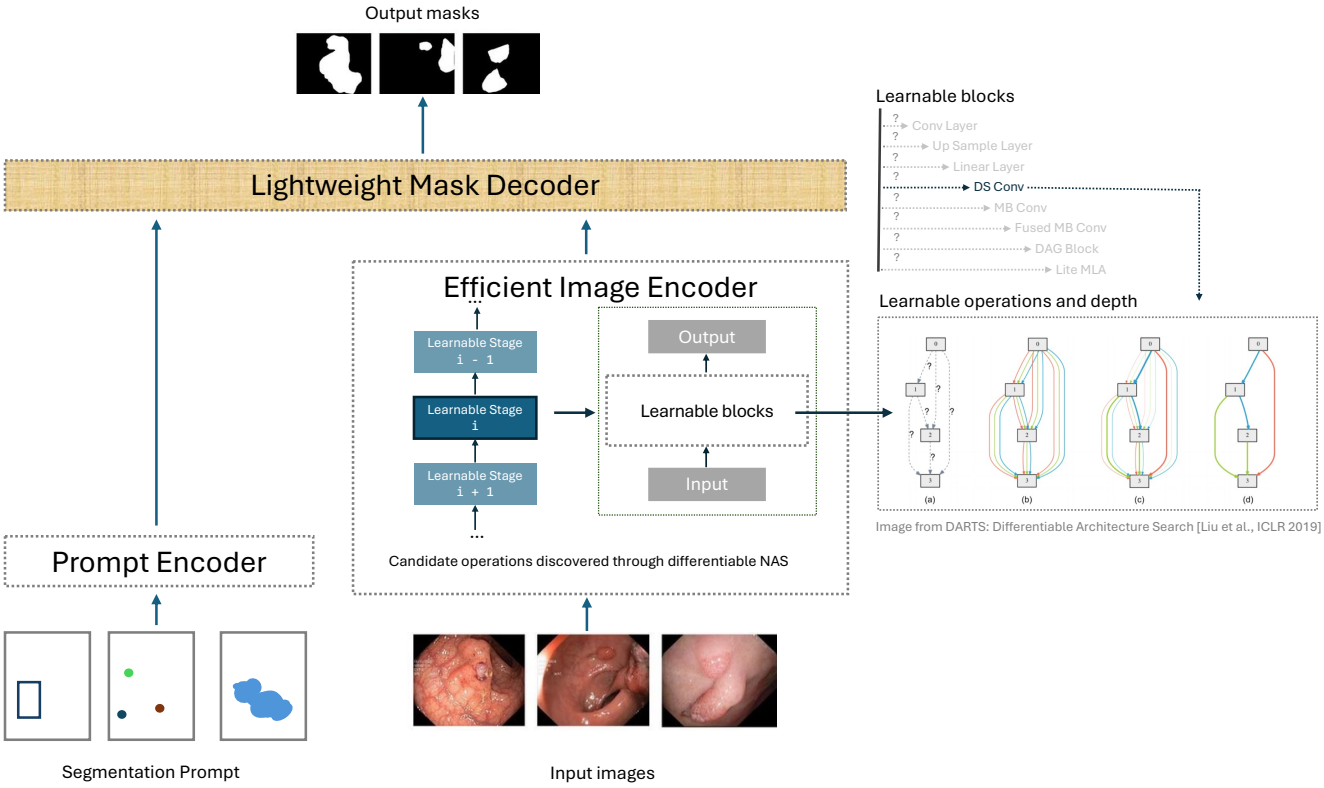

Fig. 1. Overview of the approach to obtain candidate architectures. The input image and segmentation prompts are passed through an image encoder, which consists of 5 stages that undergo neural architecture search. Each stage comprises learnable blocks with residual connections, which have been shown to be efficient in literature. The learnable blocks contain learnable operations and depths, such as ConvLayer, UpSampleLayer, LinearLayer, DepthwiseConv (DS Conv)[8], MBConv[27], FusedMBConv[29], DAGBlock[13], and LiteMLA[15]. The prompt encoder and mask decoder remain unchanged.

to achieve high segmentation accuracy while maintaining a lightweight architecture.

Our work builds upon the success of MedSAM and addresses its limitations by proposing EfficientMedSAM, a family of high-speed and memory-efficient foundation models for universal medical image segmentation. We leverage neural architecture search and knowledge distillation techniques to discover efficient transformer-based architectures that maintain high segmentation accuracy while significantly reducing computational complexity and memory footprint. By tailoring our methods to the specific requirements of medical image segmentation tasks and resource-constrained environments, we aim to bridge the gap between the performance of MedSAM and the efficiency needed for real-time clinical applications.

III. Methods

In this section, we introduce the proposed EfficientMedSAM framework, which consists of two main components: neural architecture search (NAS) for discovering efficient image encoder architectures and knowledge distillation for transferring knowledge from larger MedSAM models to the discovered efficient architectures (Figure 1).

A. Neural Architecture Search for Efficient Image Encoders

To find efficient image encoder architectures for MedSAM, we employ differentiable neural architecture search (DNAS). DNAS allows for the efficient exploration of a large search space by relaxing the discrete architecture representations into continuous ones, enabling gradient-based optimization of the architecture parameters.

B. Search Space

We define a cell-based search space for the image encoder architecture, inspired by the success of cell-based NAS in other domains. A cell is a modular building block that represents a subset of the overall network architecture. Our search space consists of various candidate operations, including convolutional and transformer-based components such as depthwise separable convolutions (DSConv), MobileNet blocks (MBConv, FusedMBConv), and efficient ViT blocks (EfficientViTBlock). By considering a diverse set of operations, we aim to discover architectures that strike a balance between efficiency and representational power. In addition to the cell architecture, the search space also includes the number of layers in the network, allowing the search algorithm to explore different combinations of cell architectures and network depths.

C. Differentiable Architecture Search

We represent the cell architecture as a directed acyclic graph (DAG), where each node represents a feature map, and each edge (i, j) represents an operation $o_{i,j}$ from the candidate set.

The operation $o_{i,j}$ transforms the feature map x_i at node i to the feature map x_j at node j:

$$x_j = \sum_{i<j} o_{i,j}(x_i) \tag{1}$$

To make the search process differentiable, we associate each edge (i, j) with a set of continuous architecture parameters $\alpha_{i,j} = \alpha^1_{i,j}, \alpha^2_{i,j}, \ldots, \alpha^N_{i,j}$, where N is the number of candidate operations. The architecture parameters determine the importance of each candidate operation:

$$\bar{o}_{i,j}(x_i) = \sum k = 1^N \frac{\exp(\alpha^k_{i,j})}{\sum_{l=1}^N \exp(\alpha^l_{i,j})} o^k_{i,j}(x_i) \tag{2}$$

During the search process, we construct an over-parameterized network that includes all candidate operations, and the output of each edge is a weighted sum of the outputs of all candidate operations, where the weights are determined by the architecture parameters α. This allows for the joint optimization of the network weights w and architecture parameters α using gradient descent:

$$\min_{w,\alpha} \mathcal{L}_{seg}(w, \alpha) \tag{3}$$

The objective of the search process is to minimize the segmentation loss $\mathcal{L}seg$, which we compute using a combination of the dice loss $\mathcal{L}dice$ and mask loss $\mathcal{L}_{mask}$, which is a binary cross entropy loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{dice} + 20 \times \mathcal{L}_{mask} \tag{4}$$

By optimizing for segmentation performance, we ensure that the discovered architectures are well-suited for the medical image segmentation task.

D. Deriving the Final Architectures

Once the search process is complete, we derive the final compact architecture by pruning the redundant operations based on the learned architecture parameters α. We select the top-k operations with the highest weights for each edge, where k is a hyperparameter that controls the trade-off between efficiency and performance:

$$o_{i,j} = \max_{o^k_{i,j}} \alpha^k_{i,j}, \quad k = 1, 2, \ldots, K \tag{5}$$

The resulting compact architectures are then trained from scratch on the target medical image segmentation task to obtain the final EfficientMedSAM models.

E. Training and Evaluation of candidate architectures

We train and evaluate the proposed EfficientMedSAM candidate architectures on the Kvasir-SEG dataset, which consists of 1000 encodscopic images with corresponding segmentation masks. The dataset covers various anatomical landmarks and pathological findings, and so it provides a suitable benchmark for medical image segmentation. We randomly split the dataset into 80% for training, 10% for validation and 10% for testing. During training, we employ standard data augmentation techniques (random cropping, flipping, rotation) to improve the model's robustness. We use a combination of dice loss and mask loss as the training objective, which helps the model learn to accurately segment the regions of interest. For evaluation, we measure the mean average precision (mAP) (see section IV for evaluation metrics) of the models on the test set of the Kvasir-SEG dataset. We also assess the computational efficiency of the models by measuring the number of Multiply-Accumulate operations (MACs) and model parameters, as well as the inference throughput on an A100 GPU (See Figure 2 and 3).

F. Knowledge Distillation

To further improve the performance of the discovered EfficientMedSAM architectures, we employ knowledge distillation to transfer knowledge from large, more complex MedSAM models. We treat the larger MedSAM model as the teacher and the EfficientMedSAM models as the students. The student model is trained to mimic the behavior of the teacher by minimizing a combination of the segmentation loss $\mathcal{L}_{seg}$ and a distillation loss $\mathcal{L}_{distill}$:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{distill} \tag{6}$$

where λ is a hyperparameter that controls the balance between the segmentation loss and the distillation loss. The distillation loss is computed as the Kullback-Leibler divergence between the output probability distributions of the student and teacher models:

$$L_{distill} = \sum i = 1^N p_T(y_i|x) \log \frac{p_T(y_i|x)}{p_S(y_i|x)} \tag{7}$$

where $p_T(y_i|x)$ and $p_S(y_i|x)$ are the output probability distributions of the teacher and student models, respectively, for the i-th pixel of the input image x. By distilling knowledge from the teacher model, the student models benefit from the learned representations and decision boundaries of the larger model while maintaining a more efficient architecture.

G. Attention Map Distillation Loss

We introduce an attention map similarity loss, which takes the cosine similarity between layers of the attention maps for the attention based student model architectures, as a proof-of-concept in transferring knowledge from attention maps from the teacher to the student model in a feature-based knowledge distillation setup. In practice and as a next step, a proxy for attention maps, such as saliency maps in the non-attention based student model architectures, is what would guide the distillation process. We further implement a novel metric that looks at global and local cosine similarities across attention maps by extracting cosine similarities from full attention maps, as well as patches of the attention maps, to measure both global and local retainment of information between the attention maps.

Definition 1 (Attention Map Similarity Loss)**.** *Let $A_T \in \mathbb{R}^{H \times W}$ and $A_S \in \mathbb{R}^{H \times W}$ be the attention maps of the teacher and student models, respectively, where H and W denote the*

height and width of the attention maps. The attention map similarity loss is defined as:

$$\mathcal{L}_{AMS} = 1 - \frac{A_T \cdot A_S}{\|A_T\|\|A_S\|} \tag{8}$$

where $\cdot$ represents the dot product and $\|\cdot\|$ denotes the Euclidean norm.

1) Global and Local Attention Map Similarity Metric:

Definition 2 (Global Attention Map Similarity)**.** *Let $A_T \in \mathbb{R}^{H\times W}$ and $A_S \in \mathbb{R}^{H\times W}$ be the attention maps of the teacher and student models, respectively. The global attention map similarity is defined as:*

$$\mathcal{S}_{global} = \frac{A_T \cdot A_S}{\|A_T\|\|A_S\|} \tag{9}$$

Definition 3 (Local Attention Map Similarity)**.** *Let $A_T \in \mathbb{R}^{H\times W}$ and $A_S \in \mathbb{R}^{H\times W}$ be the attention maps of the teacher and student models, respectively. Let $P_T^{(i,j)} \in \mathbb{R}^{h\times w}$ and $P_S^{(i,j)} \in \mathbb{R}^{h\times w}$ denote the patches of size $h \times w$ extracted from the attention maps A_T and A_S at position (i, j), respectively. The local attention map similarity is defined as:*

$$\mathcal{S}_{local} = \frac{1}{N}\sum i = 1^{H-h+1} \sum_{j=1}^{W-w+1} \frac{P_T^{(i,j)} \cdot P_S^{(i,j)}}{\|P_T^{(i,j)}\|\|P_S^{(i,j)}\|} \tag{10}$$

where $N = (H - h + 1)(W - w + 1)$ is the total number of patches.

Definition 4 (Global and Local Attention Map Similarity Metric)**.** *The global and local attention map similarity metric is defined as the weighted sum of the global and local attention map similarities:*

$$\mathcal{S}_{GL} = \alpha\mathcal{S}_{global} + (1 - \alpha)\mathcal{S}_{local} \tag{11}$$

where $\alpha \in [0, 1]$ is a hyperparameter that balances the contribution of global and local similarities.

IV. Experimental Results

A. Dataset and Evaluation metrics

To assess the segmentation performance, we employ two evaluation metrics: mean Average Precision (mAP) and Dice Loss. While Dice Loss is commonly used for evaluating segmentation tasks, we adopt mAP to evaluate the performance of the EfficientMedSAM variants, as shown in Table II. To compute mAP, we first generate bounding boxes from the ground truth segmentation masks by finding the minimum and maximum coordinates of the pixels belonging to each object. These bounding boxes are then used as prompts for the EfficientMedSAM models, which generate segmentation masks for the objects within the provided bounding boxes (See Figure 2 for performance).

The predicted segmentation masks are compared against the ground truth masks within the corresponding bounding boxes using the mAP metric. mAP measures the average precision across different recall values. By using bounding boxes as prompts, we evaluate the model's ability to accurately segment objects given their localization information. Dice Loss, on the other hand, is used to evaluate the knowledge distillation results and ablation study, as presented in Tables III and IV. Dice Loss quantifies the dissimilarity between the predicted and ground truth segmentation masks at the pixel level, with lower values indicating better segmentation performance.

B. EfficientMedSAM Variants

We present three variants of EfficientMedSAM: `efficientmedsam_l0`, `efficientmedsam_l1`, and `efficientmedsam_l2`. These variants differ in their model size and computational complexity. Knowledge distillation from the larger teacher model is performed here.

TABLE I
ARCHITECTURAL DETAILS OF EFFICIENTMEDSAM VARIANTS.

Model	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
EfficientMedSAM 0	1 x 32	1 x 64	1 x 128	4 x 256	4 x 512
EfficientMedSAM 1	1 x 32	1 x 64	1 x 128	6 x 256	6 x 512
EfficientMedSAM 2	1 x 32	2 x 64	2 x 128	8 x 256	8 x 512

The represent the different stages of the backbone network. Each cell in these columns shows the number of layers and the width (number of channels) for that stage. For example,"1 x 32" means there is 1 layer with a width of 32 channels.

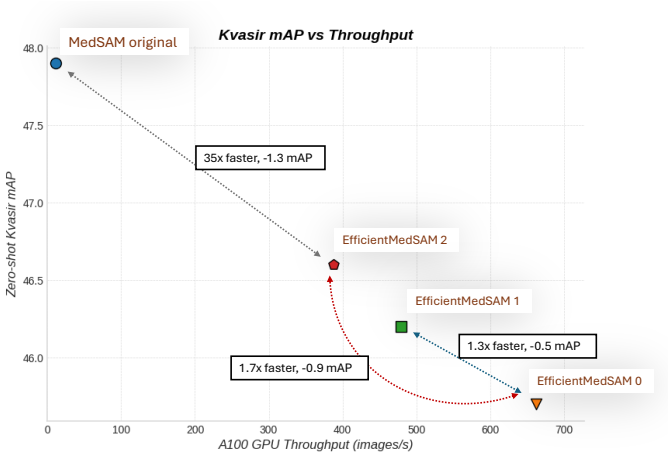

Fig. 2. Comparison between MedSAM and EfficientMedSAM variants. It's important to note that while mAP is typically used for object detection tasks, its use in this segmentation problem is justified by the adopted evaluation approach. By generating bounding boxes from the segmentation masks and using them as prompts, we assess the model's performance in accurately segmenting objects within the provided regions, similar to an object detection evaluation pipeline.

TABLE II
PERFORMANCE AND EFFICIENCY METRICS OF EFFICIENTMEDSAM VARIANTS ON THE KVASIR DATASET.

Model	mAP	Params	MACs	A100 Throughput
MedSAM (original)	49.07	641M	2973G	11 images/s
EfficientMedSAM 0	45.65	34.8M	35G	662 images/s
EfficientMedSAM 1	46.19	47.7M	49G	479 images/s
EfficientMedSAM 2	46.62	61.3M	69G	388 images/s

As shown in Table II, the discovered non-attention based EfficientMedSAM variants achieve competitive mAP scores on the Kvasir dataset while maintaining a relatively low number of parameters and MACs. The A100 throughput indicates the efficiency of the models, with `efficientmedsam_l0` achieving the highest throughput of 662 images per second.

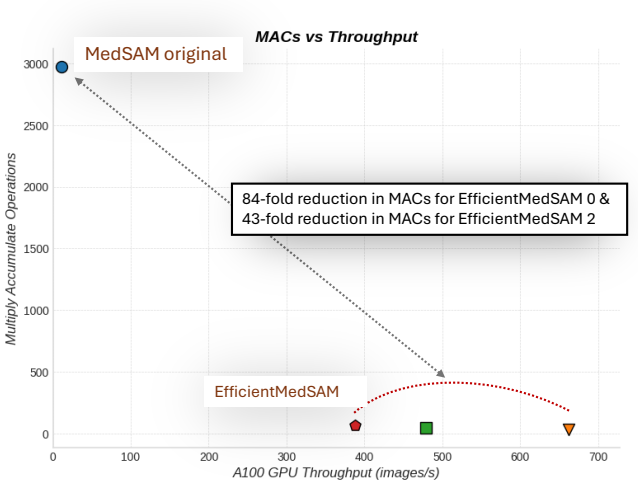

Fig. 3. Comparison of the computational efficiency between the baseline MedSAM model and the proposed EfficientMedSAM variants. The EfficientMedSAM models achieve a significant reduction in the number of multiply-accumulate operations (MACs) while maintaining high throughput, enabling more efficient medical image segmentation.

C. Knowledge Distillation Results

To further improve the performance of the EfficientMedSAM models, we apply knowledge distillation using a large teacher model. Table III presents the results of the knowledge distillation experiments on the Kvasir-SEG dataset. The student models in these experiments are based on the transformer architecture, which allows for attention-guided knowledge distillation.

TABLE III
KNOWLEDGE DISTILLATION RESULTS ON THE KVASIR DATASET.

Model	params (M)	Throughput	Dice Loss
Teacher	348.45	45.55	0.0506
Student (no attn & no logits)	97.73	102.91	0.0918
Student (only attn)	97.73	111.20	0.1343
Student (only logit)	97.73	112.77	0.1017
Student (attn & logit Loss)	97.73	100.44	0.0969

The results in Table III demonstrate the effectiveness of knowledge distillation in improving the performance of the EfficientMedSAM models. The distilled models achieve significant reductions in parameters compared to the large teacher model while maintaining low latency and high throughput. The distilled model with attention and logit loss achieves the best trade-off between dice loss and efficiency, with a dice loss of 0.0969 and an average throughput of 100.44 images per second.

D. Ablation Study

We conduct an ablation study to investigate the impact of different components in the EfficientMedSAM framework. Table III presents the results of the ablation experiments.

TABLE IV
ABLATION STUDY RESULTS ON THE KVASIR DATASET.

Model	Dice Loss	params	Latency	Throughput
EfficientMedSAM	0.0969	97.73M	0.0009s	100.44
w/o DNAS	0.1285	97.73M	0.0009s	100.44
w/o KD	0.1134	112.56M	0.0011s	90.91
w/o DNAS & KD	0.1407	112.56M	0.0011s	90.91

V. DISCUSSION

The EfficientMedSAM framework presented in this paper tackles the critical challenge of achieving accurate and efficient medical image segmentation. By leveraging differentiable neural architecture search (DNAS) and knowledge distillation, our approach discovers lightweight architectures that significantly reduce computational complexity and memory footprint while maintaining high segmentation performance. The key innovation of EfficientMedSAM lies in the use of DNAS to explore a search space consisting of efficient operations, such as depthwise separable convolutions (DSConv), MobileNet blocks (MBConv, FusedMBConv), and efficient ViT blocks (EfficientViTBlock). These operations are specifically designed to reduce computational overhead while preserving representational power. The discovered architectures, EfficientMedSAM 0 (`efficientmedsam_l0`), EfficientMedSAM 1 (`efficientmedsam_l1`), and EfficientMedSAM 2 (`efficientmedsam_l2`), exhibit a remarkable trade-off between segmentation accuracy and efficiency.

As shown in Table II, `efficientmedsam_l0` achieves a kvasir mAP of 45.65 with only 34.8M parameters and 35G MACs, resulting in an impressive throughput of 662 images/s on an A100 GPU. This represents a significant reduction in model size and computational complexity compared to state-of-the-art models like MedSAM. Notably, the slight decrease in mAP from MedSAM's 49.07 to `efficientmedsam_l0`'s 45.65 is a small trade-off for the substantial gains in efficiency. This indeed, not only shows the effectiveness of DNAS in discovering smaller and efficient architectures tailored to the specific requirements of medical image segmentation, but also, that when coupled knowledge distillation the discovered architectures can rival state-of-the-art architectures in performance.

We also investigate the effect of doing attention-guided knowledge distillation on EfficientMedSAM variants that are still heavily transformer-based. Doing attention-guided knowledge distillation from the attention-based teacher model into the non-attention-based student architectures proves to be challenging. As shown in Table IV, the distilled models (which still are transformer-based) are also still able to have a significant reduction in parameters, from 348.45 M in the teacher model to 97.73 M in the transformer-based student models. Despite this compression, the best-performing distilled model,

with both attention and logit loss, still experiences a test dice loss of 0.0969, higher than the teacher model's 0.0506.

The ablation study in Table IV highlights the individual contributions of DNAS and knowledge distillation. Removing DNAS leads to a 32.6% increase in dice loss, while removing knowledge distillation results in a 17.1% increase. This emphasizes the complementary nature of these components in achieving the best trade-off between accuracy and efficiency.

Despite the challenges in knowledge distillation, the EfficientMedSAM framework still achieves remarkable efficiency gains. The distilled model with attention and logit loss shows an average latency of 0.0009 seconds and an average throughput of 100.44 images/second, a 2.2× speedup compared to the teacher model. This demonstrates the potential of our approach in enabling real-time medical image segmentation on resource-constrained devices.

However, there is still room for improvement in the knowledge distillation process. Future research could explore more advanced techniques, such as attention transfer and feature-based distillation, to better capture the knowledge from attention-based teacher models and transfer it to the learned non-attention-based student architectures. Moreover, the current study focuses on the Kvasir dataset, which consists of endoscopic images. Further evaluation on a wider range of medical image modalities and datasets would help assess the generalizability of the EfficientMedSAM framework.

VI. Conclusion

In this paper, we presented EfficientMedSAM, a set of fast and efficient frameworks for accurate medical image segmentation. By leveraging differentiable neural architecture search (DNAS) and knowledge distillation, our approach discovers lightweight and high-performing architectures for the image encoder component of the MedSAM model. Experimental results on the challenging Kvasir dataset demonstrate that EfficientMedSAM achieves competitive segmentation accuracy compared to state-of-the-art MedSAM models while significantly reducing computational complexity and memory footprint. We believe that EfficientMedSAM serves as a proof-of-concept and paves the way for large-scale distributed training on more extensive datasets, such as SA-Med2D-20M, to further advance efficient medical image segmentation.

Acknowledgment

We thank Dr. Hadjer Benzamaine, Professor Kaoutar El Maghroui for their mentorship, and Professor Carl Vondrick for providing us with the compute to run experiments along with the I.I. Rabi Scholars Program at Columbia University.

References

[1] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. *An Analysis of Deep Neural Network Models for Practical Applications*. 2017. arXiv: 1605.07678 [cs.CV].

[2] Guobin Chen et al. "Learning Efficient Object Detection Models with Knowledge Distillation". In: *Neural Information Processing Systems*. 2017. URL: https://api.semanticscholar.org/CorpusID:29308926.

[3] Guobin Chen et al. "Learning Efficient Object Detection Models with Knowledge Distillation". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/e1e32e235eee1f970470a3a6658dfdd5-Paper.pdf.

[4] Rewon Child et al. *Generating Long Sequences with Sparse Transformers*. 2019. arXiv: 1904.10509 [cs.LG].

[5] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[6] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. *Neural Architecture Search: A Survey*. 2019. arXiv: 1808.05377 [stat.ML].

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[8] Andrew G. Howard et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv: 1704.04861 [cs.CV].

[9] Angelos Katharopoulos et al. *Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*. 2020. arXiv: 2006.16236 [cs.LG].

[10] Sungwoong Kim et al. "Scalable Neural Architecture Search for 3D Medical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019, pp. 220–228. ISBN: 9783030322489. DOI: 10.1007/978-3-030-32248-9_25. URL: http://dx.doi.org/10.1007/978-3-030-32248-9_25.

[11] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV].

[12] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. *Reformer: The Efficient Transformer*. 2020. arXiv: 2001.04451 [cs.LG].

[13] Guohao Li and Hao Zhang. "DAG-CNN: A Directed Acyclic Graph Convolutional Neural Network". In: *IEEE International Conference on Image Processing*. 2019, pp. 2716–2720.

[14] Geert Litjens et al. "A Survey on Deep Learning in Medical Image Analysis". In: *CoRR* abs/1702.05747 (2017). arXiv: 1702.05747. URL: http://arxiv.org/abs/1702.05747.

[15] Dongze Liu et al. "Lite Vision Transformer with Enhanced Self-Attention". In: *arXiv preprint arXiv:2205.14213* (2022).

[16] Hanxiao Liu, Karen Simonyan, and Yiming Yang. *DARTS: Differentiable Architecture Search*. 2019. arXiv: 1806.09055 [cs.LG].

[17] Yifan Liu et al. “Structured Knowledge Distillation for Semantic Segmentation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2599–2608. DOI: 10.1109/CVPR.2019.00271.

[18] Jun Ma et al. “Segment anything in medical images”. In: *Nature Communications* 15.1 (Jan. 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-44824-z. URL: http://dx.doi.org/10.1038/s41467-024-44824-z.

[19] Paul Michel, Omer Levy, and Graham Neubig. *Are Sixteen Heads Really Better than One?* 2019. arXiv: 1905.10650 [cs.CL].

[20] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *2016 Fourth International Conference on 3D Vision (3DV)*. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[21] Dzung Pham, Chenyang Xu, and Jerry L. Prince. “Current methods in medical image segmentation”. English (US). In: *Annual Review of Biomedical Engineering* 2.2000 (2000), pp. 315–337. ISSN: 1523-9829. DOI: 10.1146/annurev.bioeng.2.1.315.

[22] Konstantin Pogorelov et al. *KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection*. Association for Computing Machinery, 2017. URL: https://doi.org/10.1145/3193289.

[23] Dian Qin et al. “Efficient Medical Image Segmentation Based on Knowledge Distillation”. In: *IEEE Transactions on Medical Imaging* 40.12 (Dec. 2021), pp. 3820–3831. ISSN: 1558-254X. DOI: 10.1109/tmi.2021.3098703. URL: http://dx.doi.org/10.1109/TMI.2021.3098703.

[24] Adriana Romero et al. “FitNets: Hints for Thin Deep Nets”. In: *CoRR* abs/1412.6550 (2014). URL: https://api.semanticscholar.org/CorpusID:2723173.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[26] Saikat Roy et al. *SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model*. 2023. arXiv: 2304.05396 [eess.IV].

[27] Mark Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4510–4520.

[28] Dinggang Shen, Guorong Wu, and Heung-Il Suk. “Deep Learning in Medical Image Analysis”. In: *Annual Review of Biomedical Engineering* 19.Volume 19, 2017 (2017), pp. 221–248. ISSN: 1545-4274. DOI: https://doi.org/10.1146/annurev-bioeng-071516-044442. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-bioeng-071516-044442.

[29] Mingxing Tan and Quoc V Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *International Conference on Machine Learning*. 2019, pp. 6105–6114.

[30] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: 1905.11946 [cs.LG].

[31] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[32] Elena Voita et al. “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5797–5808. DOI: 10.18653/v1/P19-1580. URL: https://aclanthology.org/P19-1580.

[33] Sinong Wang et al. *Linformer: Self-Attention with Linear Complexity*. 2020. arXiv: 2006.04768 [cs.LG].

[34] Yu Weng et al. “NAS-Unet: Neural Architecture Search for Medical Image Segmentation”. In: *IEEE Access* 7 (2019), pp. 44247–44257. DOI: 10.1109/ACCESS.2019.2908991.

[35] Sirui Xie et al. *SNAS: Stochastic Neural Architecture Search*. 2020. arXiv: 1812.09926 [cs.LG].

[36] Jin Ye et al. *SA-Med2D-20M Dataset: Segment Anything in 2D Medical Imaging with 20 Million masks*. 2023. arXiv: 2311.11969 [eess.IV].

[37] Xiangyu Zhang et al. *ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices*. 2017. arXiv: 1707.01083 [cs.CV].

[38] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. “Segment Anything Model for Medical Image Segmentation: Current Applications and Future Directions”. In: *Computers in Biology and Medicine* 171 (2024), p. 108238.

[39] Barret Zoph and Quoc V. Le. *Neural Architecture Search with Reinforcement Learning*. 2017. arXiv: 1611.01578 [cs.LG].